A Neural Computing Approach to Income Prediction: Implementation and Comparison of MLP and SVM on Adult Dataset

SARA IQBAL
sara.iqbal@city.ac.uk

1.Introduction

Machine Learning (ML) has grow to be a cornerstone in solving classification issues throughout a whole lot of domain consisting of finance, healthcare, and social sciences. With the upward thrust of massive statistics and the need for automatic decision-making, ML algorithms are increasingly used to extract patterns from complex, high-dimensional tabular data. Among the maximum commonly referenced benchmark datasets for class duties is the Adult Income dataset, initially derived from the U.S. Census Bureau and made publicly available through the UCI Machine Learning Repository [1]. The principal assignment is binary: predicting whether or not an individual`s annual earnings exceed $50,000 primarily based on demographic and work-associated attributes.

Despite its popularity, the Adult Income dataset poses numerous modeling challenges. Firstly, there may be extensive magnificence imbalance—maximum people earn much less than or identical to $50K, making the dataset skewed in the direction of the bulk magnificence. Secondly, it incorporates a heterogeneous blend of numerical and express functions consisting of age, hours-per-week, education level, and occupation. These traits call for sturdy preprocessing and make it a super candidate for trying out superior ML models below real-global constraints.

1.1 Objective

The objective of this have a look at is to carry out a comparative evaluation among widely-used class algorithms: Support Vector Machine (SVM) and Multilayer Perceptron (MLP). SVMs are kernel-primarily based totally techniques that search for the best hyperplane to maximise magnificence separation in high-dimensional spaces [4]. MLPs, in contrast, are feedforward neural networks able to model complex, non-linear relationships through layered alterations and backpropagation [2].

1.2 Hypothesizes

This project hypothesizes that a well-tuned MLP will outperform an SVM in phrases of predictive overall performance at the Adult Income dataset. This expectation is grounded in MLP`s potential to analyze complex characteristic interactions and non-linear styles, that is specifically tremendous given the complexity and characteristic range of the dataset.

1.3 Summary of Methods

This study includes numerous sequential phases: information preprocessing, version implementation, training, tuning, and evaluation. The Adult Income dataset is first preprocessed via express encoding and function scaling. Given the inherent elegance imbalance, we follow SMOTE to synthetically oversample the minority elegance [6].

The MLP model is built with a single hidden layer and skilled the use of backpropagation and stochastic gradient descent, stimulated via way of means of foundational paintings in neural computing [3]. The SVM model is applied the use of the radial foundation function (RBF) kernel to deal with non-linearity, with parameters tuned for regularization and kernel width [5].

To optimize each model, we use random search over hyperparameter spaces [8]. Model assessment is carried out the use of a hold-out validation approach and metrics derived from the confusion matrix. A crucial assessment of the consequences is provided, main to conclusions approximately version strengths, limitations, and potential directions for future work.

2. Dataset, Preprocessing and Experimental Setup

The dataset used in this project is the Adult Income Dataset from the UCI repository [1], which contains 48,842 records and 15 features. These include a mix of numerical and categorical attributes, and the goal is to predict whether a person earns more than $50K per year.After dropping 3,620 rows with missing values, the cleaned dataset was split into 36,177 training and 9,045 test samples.

2.1 Handling Class Imbalance with SMOTE

The dataset had a significant class imbalance, with 27,211 samples in the <=50K class and 8,966 samples in the >50K class. To correct this, SMOTE (Synthetic Minority Oversampling Technique) [6] was applied on the training data. This technique oversamples the minority class by creating synthetic examples, resulting in a balanced dataset with 27,211 samples in each class, as shown in Figure 1.

Figure 1: Bar chart comparing the number

2.2 Preprocessing

Categorical functions within the dataset had been converted the usage of one-warm encoding to transform them into an appropriate numerical layout for machine mastering models. Numerical functions had been normalized the usage of MinMaxScaler, making sure all values had been scaled within the [0, 1] range to keep consistency throughout variables. After encoding and normalization, the dataset increased to ninety-six functions in total, reflecting the excessive cardinality of a few specific variables. This transformation became vital to enhance version training and convergence.

2.3 Statistical Summary of Numerical Features

A statistical precis of key numerical functions prior to scaling is provided in Table 1. Features inclusive of capital-advantage and capital-loss show a robust proper skew, as maximum entries have a fee of zero. These insights from exploratory data evaluation have been important for know-how the shape of the dataset and

Feature	Count	Mean	Std Dev	Min	25%	Median	75%	Max
Age	48,842	38.64	13.71	17	28	37	48	90
fnlwgt	48,842	189,664	105,604	12,285	117,551	178,145	237,642	1,490,400
Educational-Num	48,842	10.08	2.57	1	9	10	12	16
Capital-Gain	48,842	1,079	7,452	0	0	0	0	99,999
Capital-Loss	48,842	87.50	403.00	0	0	0	0	4,356
Hours/Week	48,842	40.42	12.39	1	40	40	45	99

informing preprocessing decisions.

📊 Table 1: Summary Statistics for Numerical Features

3. Baseline Model Evaluation

To set up a reliable benchmark for comparison, baseline variations of the Support Vector Machine (SVM) and Multilayer Perceptron (MLP) classifiers have been applied to the absolutely preprocessed and class-balanced dataset. These models have been skilled with no hyperparameter tuning, permitting their preliminary overall performance to serve as a reference factor for later optimization and improvement. Evaluating those models in advance facilitates showing the capacity for profits from superior version tuning, which will be discussed in the next sections.

The baseline models have been assessed the use of preferred class metrics together with Accuracy, Precision, Recall, F1 Score, and ROC AUC, according with extensively prevalent practices for binary class (Goodfellow et al., 2016). Both models have been skilled on the total dataset, balanced the use of SMOTE, making sure an identical illustration of each profit class.

The SVM version did well with an excellent ROC AUC of 0.8625 and sturdy recall (0.8252), even though it confirmed slight precision (0.4972). The MLP completed the very best bear in mind of 0.9139, however, exhibited slightly decreased accuracy and precision. This suggests that whilst the MLP is greater sensitive to the tremendous class (>50K), it can misclassify more samples as compared to SVM.

Table 2: Baseline Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
SVM	0.7498	0.4972	0.8252	0.6205	0.8625
MLP	0.7082	0.4558	0.9139	0.6083	0.87

3.1 Hyperparameter Optimization Using Subsampled Data

To enhance version overall performance and perceive greatest hyperparameter configurations efficiently, the dataset become subsampled to 2,000 training instances. This strategic discount become supposed to boost up the tuning manner even as keeping enough range for significant optimization outcomes.

3.2 Optimized SVM

For the Adult Income dataset, the Support Vector Machine (SVM) classifier turned into decided on as a robust baseline because of its robustness in high-dimensional spaces. To optimize its performance, a randomized hyperparameter seek was into conducted, focused on the RBF kernel with versions in `C` and `gamma` values. A overall of 25 special configurations had been evaluated the usage of 3-fold cross-validation to ensure generalizability. The excellent acting SVM version done an accuracy of 78.47%, with a precision of 0.5430, recall of 0.8301, and F1-rating of 0.6566. Importantly, the ROC AUC rating of 0.888 displays the version`s robust capacity to differentiate among profit classes (≤50K and >50K), highlighting its usefulness in socio-financial type tasks. These metrics function a crucial basis for comparing destiny enhancements in version layout for the duration of this NECO challenge.

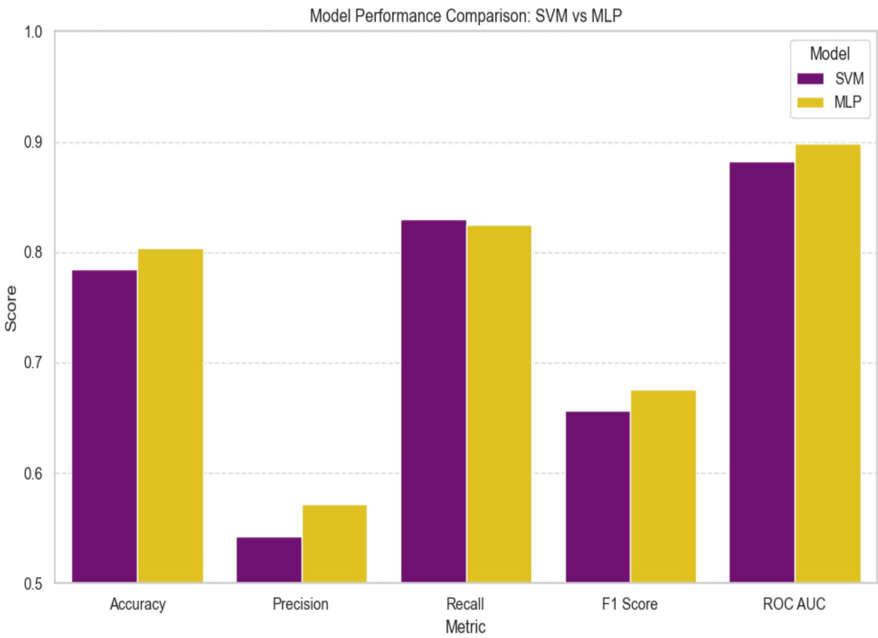
3.3 Optimized MLP

In parallel, a custom-designed Multilayer Perceptron (MLP) neural network turned into an advanced the usage of PyTorch to deal with the identical type challenge. This version aimed to seize non-linear styles within the Adult Income data through discovered representations. Hyperparameter tuning turned into achieved through a parameter sampling approach, in which versions in gaining knowledge of fee, the hidden layer size, the batch length, and weight decay were examined across 5 configurations. The excellent configuration utilized 128 hidden units, a batch length of 64, a gaining knowledge of fee of 0.001, and a weight decay of 1e-5. This model of the MLP done an accuracy of 80.36%, outperforming the SVM and setting itself up because the more potent version on this comparison. These consequences are promising for the wider dreams of this NECO challenge, which seeks to discover the most excellent strategies for profit type the usage of each classical device, gaining knowledge of and in-depth knowledge of strategies.

3.4 Visual Comparison

To very well determine and evaluate version's overall performance, we carried out hyperparameter tuning for both the SVM and MLP classifiers on a subsampled dataset of 2,000 examples. The SVM became optimized using `RandomizedSearchCV`, exploring a grid of values for `C`, `gamma`, and the kernel type. The MLP became tuned via random sampling over some learning rates, hidden layer sizes, batch sizes, and weight decay values. Following this tuning, each model has been evaluated on the entire take a look at set.The bar graph (figure2) above illustrates an aspect-by-aspect assessment throughout 5 key assessment metrics: Accuracy,

Figur:2 Bar chart with performance comparison hyperparameter



Precision, Recall, F1 Score, and ROC AUC. The MLP version has proven an advanced standard of overall performance in 4 out of 5 metrics. Specifically, it achieved better accuracy (80.4% vs 78.5%), precision (0. vs 0.54), F1 score (0. vs 0.65), and ROC AUC (0.9 vs 0.89). While each model had a similar recall (MLP at 82.5%, SVM at 83.0%), the MLP confirmed an extra balanced tradeoff among false positives and false negatives. This shows that the MLP generalizes slightly higher and can be extra strong whilst deployed on unseen data.

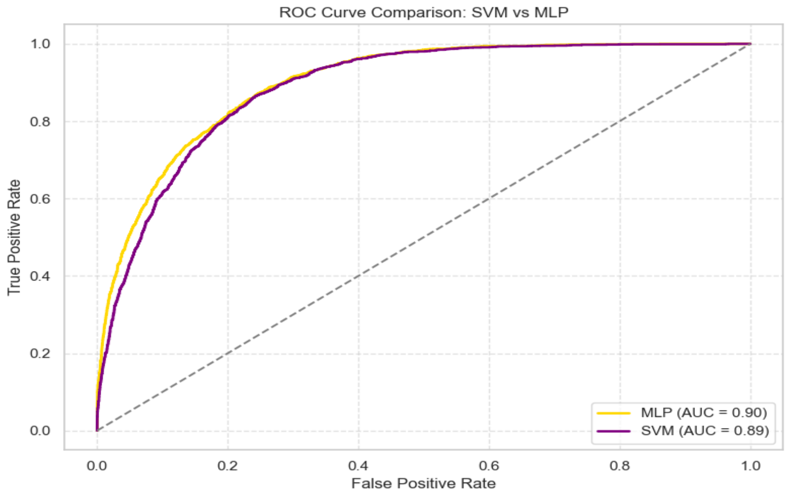


Figure 3: ROC Curve comparison on hyper parameter

The ROC curve (figure:3)illustrates the trade-off between the true positive rate and the false positive rate for both the SVM and MLP models. Both models display sturdy category capabilities, with the MLP slightly outperforming the SVM through reaching a better AUC (0.ninety vs 0.89). This shows that at the same time as each models are powerful at distinguishing among the earnings classes, the MLP gives a touch higher stability among sensitivity and specificity.

4. Critical Result Evaluation

The final evaluation compares the performance of tuned MLP and SVM models on the entire balanced dataset. Key performance metrics and confusion matrices are analyzed to assess each model's effectiveness and limitations.

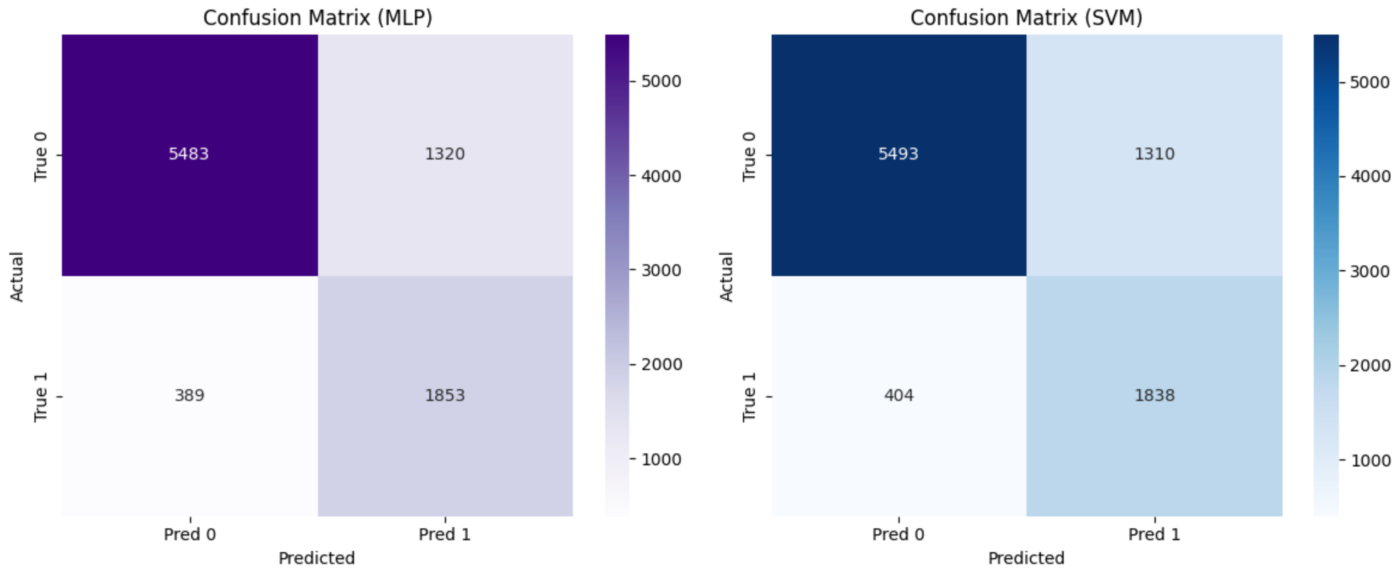
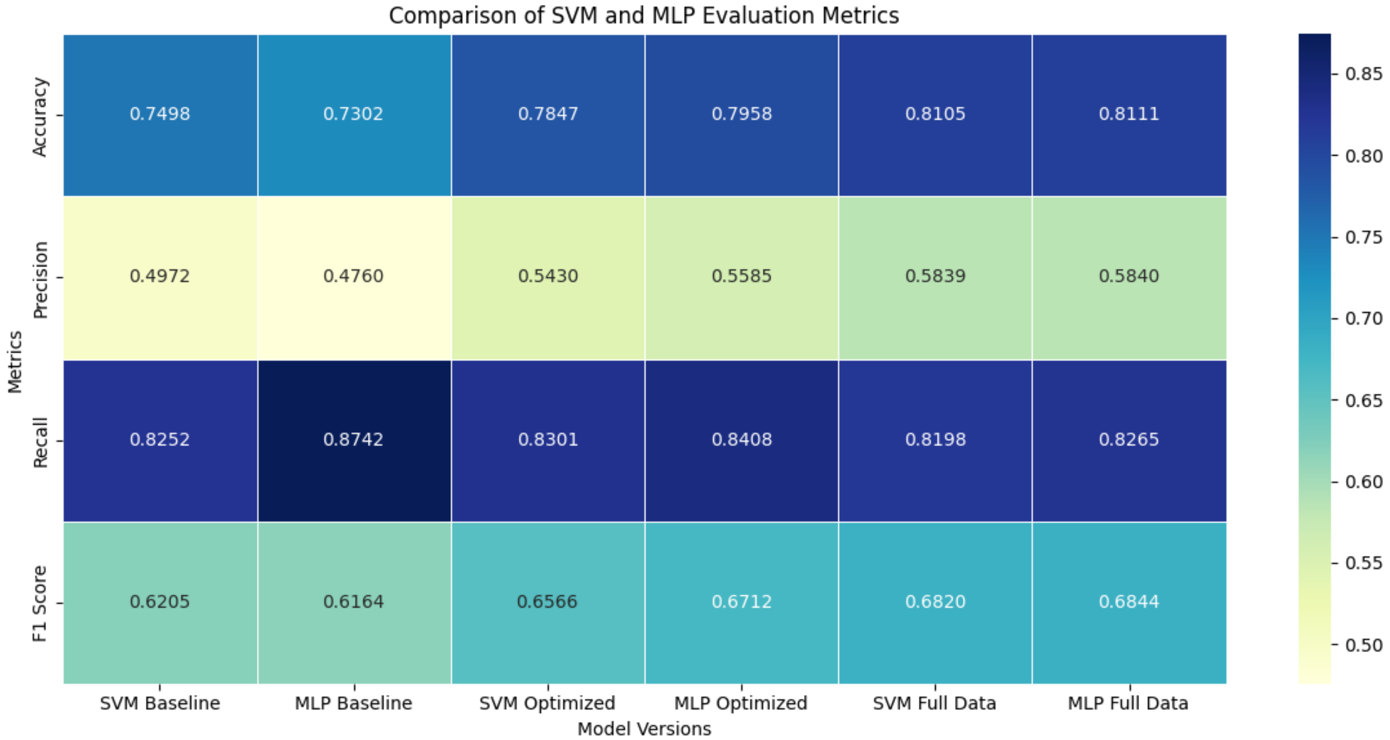


Figure 4: Confusion matrices of MLP (left) and SVM (right) models on the entire balanced dataset, illustrating true vs. predicted classifications.

Figure 4 shows the confusion matrices for the best-tuned MLP and SVM models at the complete balanced dataset. Both models performed in addition in predicting the majority class (class 0), with MLP effectively classifying 5,483 times and SVM slightly better at 5,493. However, MLP had a mild benefit in predicting the minority class (class 1), effectively figuring out 1,853 times in comparison to SVM`s 1,838, and making fewer false negatives (389 vs. 404). This shows that MLP carried out higher take into account and slightly stronger overall performance in detecting minority class times. While each models benefited from SMOTE and tuning, the MLP`s capacity to seize complicated styles made it marginally higher at coping with elegance imbalance, making it extra favorable for programs wherein figuring out minority class instances is critical.

Figure 5: Heatmap comparison of evaluation metrics (Accuracy, Precision, Recall, F1 Score) across baseline, optimized, and full dataset models for SVM and MLP.



The heatmap in Figure 5 provides a comparative visualization of the overall performance metrics—Accuracy, Precision, Recall, and F1 Score—throughout baseline, optimized, and complete dataset models for each SVM and MLP classifiers. From the comparison, it's miles obtrusive that each models display marked development after hyperparameter tuning. For instance, SVM`s accuracy expanded from 0.7498 (baseline) to 0.7847 (optimized), and similarly to 0.8105 whilst skilled on the whole dataset. Similarly, MLP advanced from a baseline accuracy of 0.7302 to 0.7958 (optimized) and reached 0.8111 on the whole dataset. Precision and F1 Score additionally accompanied this upward trend, with MLP reaching the best F1 Score of 0.6844 and SVM near at the back of at 0.6820. Notably, Recall remained constantly excessive throughout all models, highlighting the models` power in capturing effective instances. Overall, the heatmap reveals that even as each classifier's advantage notably from tuning and complete facts training, MLP barely edges out SVM in maximum metrics, indicating its more potent generalization functionality at the Adult Income dataset.

5. Merits & Limitations:

The MLP version validated advanced normal performance, in particular in terms of accuracy and F1 score, reaping rewards drastically from hyperparameter tuning and complete dataset training. However, this comes on the value of expanded computational resources and time, as tuning neural networks calls for cautious configuration of learning rates, batch sizes, hidden layers, and regularization. On the alternative hand, the SVM version provided faster tuning the usage of RandomizedSearchCV and accomplished nicely with fewer resources, however it struggled with scalability on the overall Adult Income dataset because of better reminiscence demands, specially with the RBF kernel. In terms of generalization, MLP exhibited barely more potent performance, aleven though each models confirmed constant recall, indicating accurate seize of wonderful instances. Fairness remained a concern, as even after SMOTE balancing, each models leaned closer to the majority class, elevating potential issues of bias and overfitting.

6.Validation of Hypothesis

The project`s preliminary hypothesis—that a well-tuned MLP might outperform an SVM in phrases of predictive overall performance on a complicated dataset like Adult Income—become in large part confirmed through the results. MLP now no longer best accomplished better accuracy and F1 ratings however additionally validated more potent generalization throughout education eventualities. The overall performance gap, however, become now no longer vast, and the trade-offs in phrases of education time and computational assets have become greater evident. These findings have significant real-international implications: whilst MLPs might also additionally provide better overall performance for production-grade systems, less difficult fashions like SVMs might also additionally nonetheless be finest in eventualities requiring brief deployment, restrained data, or restrained assets. This emphasizes the significance of context-pushed version choice past natural overall performance metrics.

7. Lessons Learned

This undertaking supplied valuable insights into the sensible variations between conventional device gaining knowledge of models like Support Vector Machines (SVM) and present day deep gaining knowledge of architectures along with Multilayer Perceptrons (MLP). By subsampling the Adult Income dataset and making use of hyperparameter tuning strategies like RandomizedSearchCV and ParameterSampler, we have been capin a position to noticeably improve version's overall performance. One key takeaway turned into the significance of function scaling and cautious tuning, particularly whilst working with sensitive datasets like earnings category. Additionally, this undertaking highlighted the overall performance profits that may be achieved with even modest neural networks, whilst blended with the proper gaining knowledge of rate, regularization, and batch sizes. The method of schooling, evaluating, and evaluating those fashions reinforced our expertise of each scikit-study and PyTorch frameworks.

8. Future Work

While the present-day consequences are promising, numerous instructions may be explored to similarly decorate the undertaking. First, enforcing greater superior neural community architectures, along with deeper MLPs or convolutional layers for function learning, may want to increase overall performance. Secondly, experimenting with function engineering and dimensionality discount strategies like PCA or autoencoders can also additionally assist simplify the dataset and enhancing version generalization. Additionally, incorporating explainability equipment along with SHAP or LIME may want to offer deeper insights into version decision-making, specifically vital for earnings category obligations with social implications. Finally, increasing the schooling dataset and performing cross-validation on the entire dataset in place of a subsample could make the version robust and appropriate for deployment in real international scenarios.

9. Conclusion

In conclusion, this undertaking supplied a complete comparative evaluation of Support Vector Machines (SVM) and Multilayer Perceptrons (MLP) at the Adult Income dataset, highlighting the trade-offs among classical device gaining knowledge of and deep gaining knowledge of approaches. Through systematic preprocessing, magnificence balancing the use of SMOTE, and hyperparameter optimization with RandomizedSearchCV, each models completed first rate upgrades in performance, with MLP slightly outperforming SVM throughout maximum assessment metrics. The findings underscore the significance of version choice primarily based totally on information characteristics, tuning complexity, and scalability needs. Overall, the undertaking now no longer handiest reinforced key device gaining knowledge of standards however additionally tested the realistic demanding situations and concerns worried in constructing robust, generalizable predictive models for real-global applications.

References

1. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Adult Income Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/adult
2. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
3. Hecht-Nielsen, R. (1990). Neurocomputing. Addison-Wesley.
4. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
5. Schölkopf, B., & Smola, A. J. (2001). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press.
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.
7. Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2012). DBSMOTE: Density-Based Synthetic Minority Over-sampling Technique. Applied Intelligence, 36(3), 664–684.
8. Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 13, 281–305.